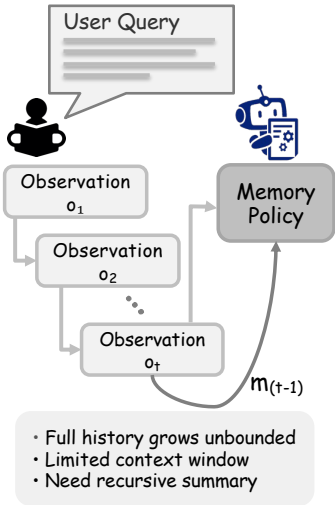


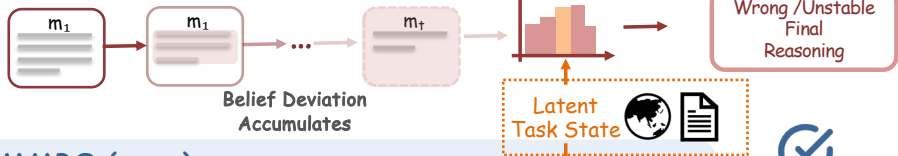
Long-Horizon Task



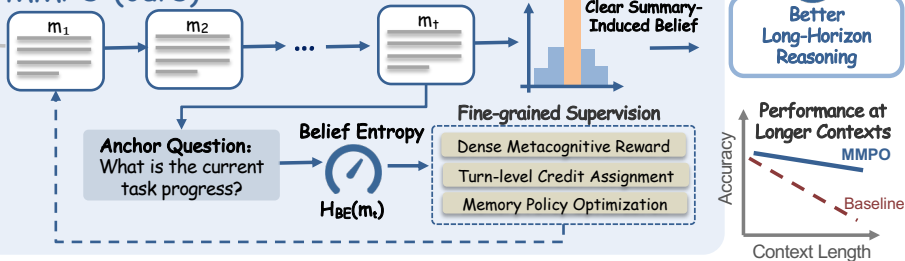
Previous Memory Agent / Outcome-Only Memory RL

Sparse terminal reward only

Coarse-grained Optimization cannot identify which intermediate summary caused noise and belief deviation



MMPO (ours)



MMPO improves long-horizon reasoning by penalizing ambiguous intermediate summaries and preserving belief clarity.